

# Pattern mining in personal demographic trajectories

Статья подготовлена в результате проведения исследования №16-05-0011 «Разработка и апробация методик анализа демографических последовательностей» в рамках Программы «Научный фонд Национального исследовательского университета «Высшая школа экономики» (НИУ ВШЭ)» в 2016-2017 гг. и в рамках государственной поддержки ведущих университетов Российской Федерации "5-100"

Dmitry I. Ignatov<sup>1</sup>, Danil Gizdatullin<sup>1</sup>, Ekaterina Mitrofanova<sup>1</sup>, Anna Muratova<sup>1</sup>, Jaume Baixeries<sup>2</sup>

<sup>1</sup>National Research University Higher School of Economics, Moscow

<sup>2</sup>Universitat Politècnica de Catalunya, Barcelona

The April Conference  
2016, Moscow

- First job (job)
- The highest education degree is obtained (education)
- Leaving parents' home (separation)
- First partner (partner)
- First marriage (marriage)
- First child birth (children)
- Break-up (parting)
- ... (divorce)

# Data and problem statement

[Ignatov et al., 2015],[Blockeel et al., 2001]

Generation and Gender Survey (GGS): three waves panel data for 11 generations of Russian citizens starting from 30s

## Binary classification

1545 men

3312 women

## Examples of sequential patterns

- $\langle \{education, separation\}, \{work\}, \{marriage\}, \{children\} \rangle (m)$
- $\langle \{work\}, \{marriage\}, \{children\} \{education\} \rangle (f)$
- $\langle \{partner\}, \{marriage, separation\}, \{children\} \rangle (f)$

# Basic definitions

Textbooks of Han et al., Zaki & Meira, Aggarwal et al., etc

- $s = \langle s_1, \dots, s_k \rangle$  is the **subsequence** of  $s' = \langle s'_1, \dots, s'_{k'} \rangle$  ( $s \preceq s'$ ) if  $k \leq k'$  and there exist  $1 \leq r_1 < r_2 < \dots < r_k \leq k'$  such  $s_j = s'_{r_j}$  for all  $1 \leq j \leq k$ .
- $support(s, D)$  is the **support** of a sequence  $s$  in  $D$ , i.e. the number of sequences in  $D$  such that  $s$  is their subsequence.

$$support(s, D) = |\{s' | s' \in D, s \preceq s'\}|$$

- $s$  is a **frequent closed sequence (sequential pattern)** if there is no  $s'$  such that  $s \prec s'$  and

$$support(s, D) = support(s', D)$$

# Example

Let  $D$  be a set of sequences:

Таблица: Dataset  $D$ .

$s_1$	$\{a, b, c\}\{a, b\}\{b\}$
$s_2$	$\{a\}\{a, c\}\{a\}$
$s_3$	$\{a, b\}\{b, c\}$

- $I = \{a, b, c\}$  is the set of all items (atomic events)
- $\langle\{a, b\}\{b\}\rangle$  belongs to  $s_1$  and  $s_3$  but it is missing in  $s_2$
- $support_D(\langle\{a, b\}\{b\}\rangle) = 2$
- $\{\langle\{a\}\rangle, \langle\{c\}\rangle, \langle\{a\}\{c\}\rangle, \langle\{a, b\}\{b\}\rangle, \langle\{a, c\}\{a\}\rangle\}$  is the set of closed sequences.

# Contiguous prefix-based sequential patterns

- $s = \langle s_1, \dots, s_k \rangle$  is a **contiguous prefix-based subsequence** of  $s' = \langle s'_1, \dots, s'_{k'} \rangle$  ( $s * = s'$ ) if  $k \leq k'$  and  $\forall i \in k' : s_i = s'_i$ .
- **Support of contiguous prefix-based sequences**  
Let  $T$  be a set of sequences.

$$\text{support}(s, T) = \frac{|\{s' | s' \in T, s * = s'\}|}{|T|}$$

# Contiguous prefix-based sequential patterns

- Let  $0 < minSup \leq 1$  be a minimal support parameter and  $D$  is a set of sequences then **searching for prefix-based contiguous sequential patterns** is the task of enumeration of all prefix-based contiguous sequences  $s$  such that  $support(s, D) \geq minSup$ . Every sequence  $s$  with  $support(s, D) \geq minSup$  is called a **prefix-based contiguous sequential pattern**.
- Prefix-based contiguous sequential pattern (PGSP)  $p$  is called **closed** if there is no PGSP  $d$  of greater or equal support such that  $d = p*$ .

## Example

Таблица:  $D$  is a set of sequences.

$s_1$	$\{a\}\{b\}\{d\}$
$s_2$	$\{a\}\{b\}\{c\}$
$s_3$	$\{a, b\}\{b, c\}$

$$s = \langle \{a\}\{b\} \rangle$$

- $I = \{a, b, c\}$  is the set of all items (atomic events)
- $s_1 = s^*$ ;  $s_2 = s^*$
- $s_3 \neq s^*$
- $Supp_D(s) = \frac{2}{3}$
- $\langle \{a\}\{b\} \rangle$  is closed,  $\langle \{a\} \rangle$  is not closed.



## Growth Rate

$$\text{growth\_rate}_{D' \rightarrow D''}(X) = \begin{cases} \frac{\text{supp}_{D''}(X)}{\text{supp}_{D'}(X)} & \text{if } \text{supp}_{D'}(X) \neq 0 \\ 0 & \text{if } \text{supp}_{D''}(X) = \text{supp}(X) = 0 \\ \infty & \text{if } \text{supp}_{D''}(X) \neq 0 \text{ and } \text{supp}_{D'}(X) = 0 \end{cases}$$

## Class score

$$\text{score}(s, C) = \sum_{e \subseteq s, e \in E(c)} \frac{\text{growth\_rate}_C(e)}{\text{growth\_rate}_C(e) + 1} \cdot \text{supp}_C(e)$$

# Emerging patterns for classification

$s$  is a new object

$$\text{normal\_score}_{\oplus}(s) = \frac{\sum_{p \in P_{\oplus}: p \sqsubseteq s} \text{GrowthRate}(p, \mathbb{K}_{\oplus}, \mathbb{K}_{\ominus})}{\text{median}(\text{GrowthRate}(P_{\oplus}))}$$

$$\text{normal\_score}_{\ominus}(s) = \frac{\sum_{p \in P_{\ominus}: p \sqsubseteq s} \text{GrowthRate}(p, \mathbb{K}_{\ominus}, \mathbb{K}_{\oplus})}{\text{median}(\text{GrowthRate}(P_{\ominus}))}$$

Classification via emerging patterns

$$\text{class}(s) = \begin{cases} \text{positive} & \text{if } \text{normal\_score}_{\oplus}(s) > \text{normal\_score}_{\ominus}(s) \\ \text{negative} & \text{if } \text{normal\_score}_{\oplus}(s) < \text{normal\_score}_{\ominus}(s) \\ \text{undetermined} & \text{if } \text{normal\_score}_{\oplus}(s) = \text{normal\_score}_{\ominus}(s) \end{cases}$$

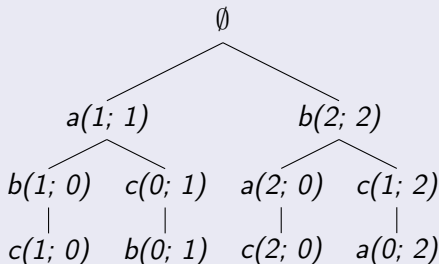
# Execution example

## Input sequences

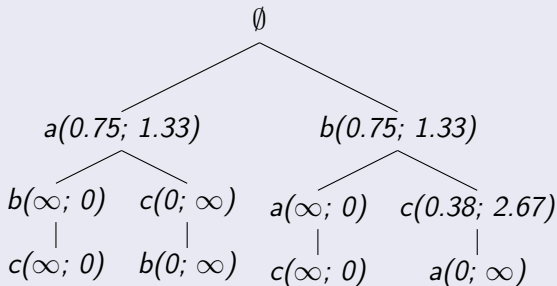
*class\_0* : {⟨{a}{b}{c}⟩, ⟨{b}{a}{c}⟩, ⟨{b}{a}{c}⟩, ⟨{b}{c}⟩}

*class\_1* : {⟨{a}{c}{b}⟩, ⟨{b}{c}{a}⟩, ⟨{b}{c}{a}⟩}

## Prefix tree



## Counting Growth Rate

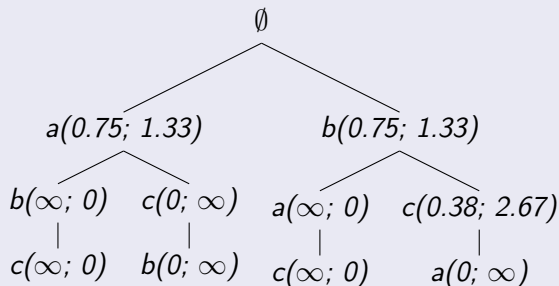


## Growth rate

$$0.75 = \frac{1}{4} / \frac{1}{3}; 1.33 = \frac{1}{3} / \frac{1}{4}$$

$$0.38 = \frac{1}{4} / \frac{2}{3}; 2.67 = \frac{2}{3} / \frac{1}{4}$$

## Computing Score



## New sequence

$$\text{minGR} = 2$$

$$\langle \{b\}; \{c\}; \{a\} \rangle - ???$$

$$\text{Score}_0 = 0$$

$$\text{Score}_1 = 2.67 + \infty = \infty$$

# Comparison of closed and non-closed patterns

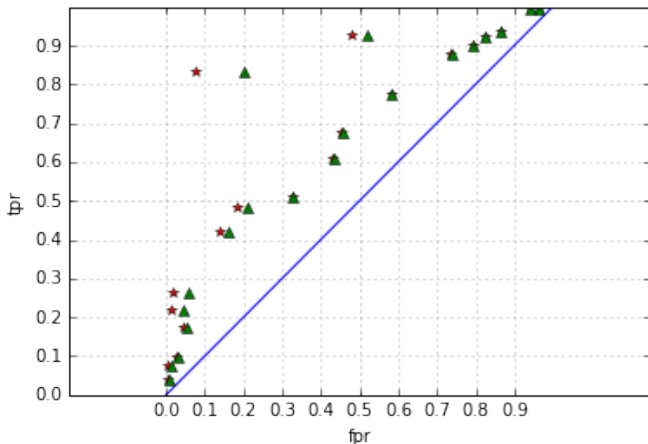


Рис.: TPR vs FPR for closed and non-closed patterns

# Experiments and results

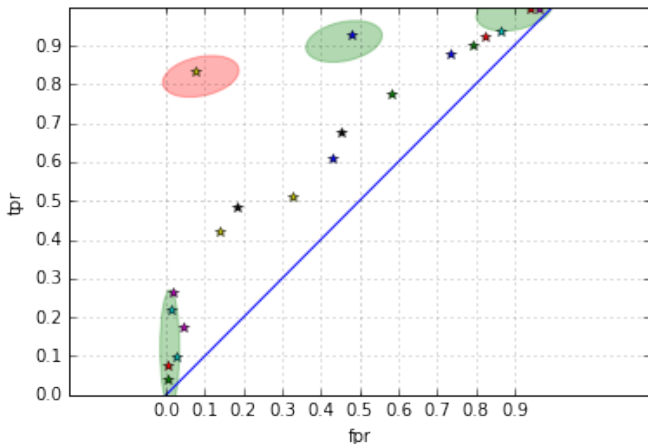


Рис.: TPR-FPR for classification by gender via contiguous prefix-based patterns

## Interesting patterns (women)

$(\langle\{work, separation\}, \{marriage\}, \{children\}, \{education\}\rangle, [\infty, 0.006])$

$(\langle\{separation, partner\}, \{marriage\}\rangle, [\infty, 0.006])$

$(\langle\{work, separation\}, \{marriage\}, \{children\}\rangle, [\infty, 0.008])$

$(\langle\{work, separation\}, \{marriage\}\rangle, [\infty, 0.009])$



## Interesting patterns (men)

$(\langle\{\textit{education}\}, \{\textit{marriage}\}, \{\textit{work}\}, \{\textit{children}\}, \{\textit{separation}\}\rangle, [10.6, 0.006])$

$(\langle\{\textit{education}\}, \{\textit{marriage}\}, \{\textit{work}\}, \{\textit{children}\}\rangle, [12.7, 0.007])$

$(\langle\{\textit{educ}\}, \{\textit{work}\}, \{\textit{part}\}, \{\textit{mar}\}, \{\textit{sep}\}, \{\textit{ch}\}\rangle, [10.6, 0.006])$

# Experiments and results

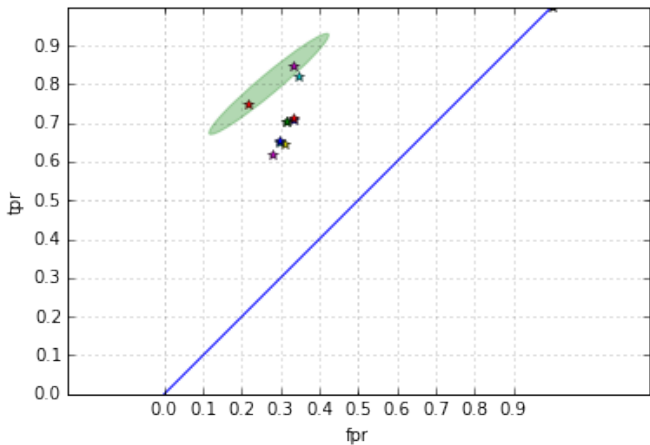


Рис.: TPR-FPR for classification by generation via contiguous prefix-based patterns

# Interesting patterns (Different Generations; Women)

## Old women

$(\langle\{work\}, \{separation\}\rangle, [1.85, 0.38])$

$(\langle\{work\}, \{marriage, separation\}\rangle, [3.92, 0.08])$

## Young women

$(\langle\{education\}\rangle, [1.84, 0.26])$

$(\langle\{education\}, \{work\}\rangle, [4.01, 0.1])$

# Experiments and results

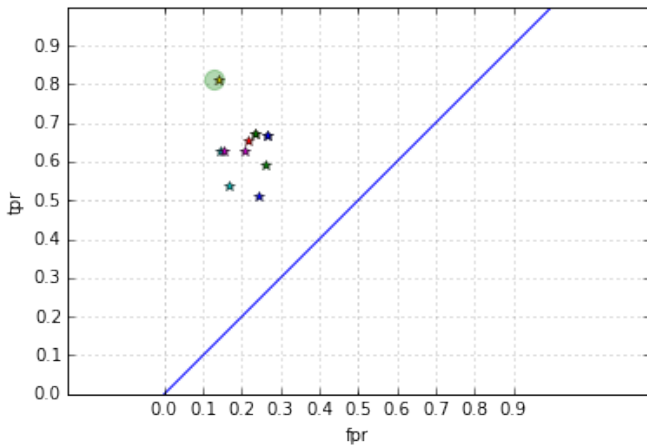


Рис.: TPR-FPR for classification by generation via contiguous prefix-based patterns

# Interesting patterns (Different Generations; Men)

## Old men

$(\langle\{\textit{work}\}, \{\textit{marriage, separation}\}, \{\textit{education}\}\rangle, [13.52, 0.025])$

$(\langle\{\textit{work}\}, \{\textit{marriage}\}, \{\textit{separation}\}\rangle, [22.87, 0.042])$

$(\langle\{\textit{work}\}, \{\textit{marriage}\}, \{\textit{separation}\}, \{\textit{education}\}\rangle, [\infty, 0.0208])$

## Young men

$(\langle\{\textit{education}\}, \{\textit{work}\}, \{\textit{separation}\}, \{\textit{marriage}\}, \{\textit{children}\}\rangle, [10.58, 0.020])$

$(\langle\{\textit{education}\}, \{\textit{work}\}, \{\textit{separation, partner}\}, \{\textit{marriage}\}\rangle, [8.65, 0.016])$

$(\langle\{\textit{education}\}, \{\textit{marriage, separation}\}\rangle, [7.69, 0.015])$

- ① We have studied several pattern mining techniques for demographic sequences including pattern-based classification in particular.
- ② We have fitted existing approaches for sequence mining of a special type (contiguous and prefix-based ones).
- ③ The results for different demographic groups (classes) have been obtained and interpreted.
- ④ In particular, a classifier based on emerging sequences and pattern structures has been proposed.

Thank you!

Questions?